

2018

A Simple Model for Cell Type Recognition Using 2D-Correlation Analysis of FTIR Images From Breast Cancer Tissue

Mohamed H.M. Ali

Qatar Biomedical Research Institute, Doha, Qatar

Fazle Rakib

Qatar University, Department of Chemistry and Earth Sciences, Doha, Qatar

Khalid A. Al-Saad

Qatar University, Department of Chemistry and Earth Sciences, Doha, Qatar


Rafif Al-Saady

Al-Ahli Hospital, Department of Pathology and Laboratory Medicine, Doha, Qatar

Fiona Lyng

Technological University Dublin, Fiona.lyng@tudublin.ie

Follow this and additional works at: <https://arrow.tudublin.ie/radart>

 Part of the [Medical Sciences Commons](#)
See next page for additional authors

Recommended Citation

Ali, M.H., Rakib, F. & Al-Saad, K. (2018). A simple model for cell type recognition using 2D-correlation analysis of FTIR images from breast cancer tissue. *Journal of Molecular Structure*, vol. 1163, pg. 472-479. doi:10.1016/j.molstruc.2018.03.044

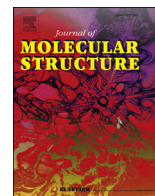
This Article is brought to you for free and open access by the Radiation and Environmental Science Centre at ARROW@TU Dublin. It has been accepted for inclusion in Articles by an authorized administrator of ARROW@TU Dublin. For more information, please contact yvonne.desmond@tudublin.ie, arrow.admin@tudublin.ie, brian.widdis@tudublin.ie.



This work is licensed under a [Creative Commons Attribution-NonCommercial-Share Alike 3.0 License](#)

Authors

Mohamed H.M. Ali, Fazle Rakib, Khalid A. Al-Saad, Rafif Al-Saady, Fiona Lyng, and Erik Goormaghtigh



A simple model for cell type recognition using 2D-correlation analysis of FTIR images from breast cancer tissue

Mohamed H. Ali ^a, Fazle Rakib ^b, Khalid Al-Saad ^b, Raffi Al-Saady ^c, Fiona M. Lyng ^d, Erik Goormaghtigh ^{e,*}

^a Qatar Biomedical Research Institute (QBRI), Hamad Bin Khalifa University (HBKU), Doha, Qatar

^b Department of Chemistry and Earth Sciences, Qatar University, Doha, Qatar

^c Pathology and Laboratory Medicine, Al Ahli Hospital, Doha, Qatar

^d DIT Centre for Radiation and Environmental Science, FOCAS Research Institute, Dublin Institute of Technology, Ireland

^e Center for Structural Biology and Bioinformatics, Laboratory for the Structure and Function of Biological Membranes, Campus Plaine CP206/02, Université Libre de Bruxelles CP206/2, B1050, Brussels, Belgium

ARTICLE INFO

Article history:

Available online 13 March 2018

Keywords:

Infrared imaging
Spectral histopathology
Infrared spectroscopy
2D correlation spectroscopy
Breast cancer

ABSTRACT

Breast cancer is the second most common cancer after lung cancer. So far, in clinical practice, most cancer parameters originating from histopathology rely on the visualization by a pathologist of microscopic structures observed in stained tissue sections, including immunohistochemistry markers. Fourier transform infrared spectroscopy (FTIR) spectroscopy provides a biochemical fingerprint of a biopsy sample and, together with advanced data analysis techniques, can accurately classify cell types. Yet, one of the challenges when dealing with FTIR imaging is the slow recording of the data. One cm² tissue section requires several hours of image recording. We show in the present paper that 2D covariance analysis singles out only a few wavenumbers where both variance and covariance are large. Simple models could be built using 4 wavenumbers to identify the 4 main cell types present in breast cancer tissue sections. Decision trees provide particularly simple models to reach discrimination between the 4 cell types. The robustness of these simple decision-tree models were challenged with FTIR spectral data obtained using different recording conditions. One test set was recorded by transfection on tissue sections in the presence of paraffin while the training set was obtained on dewaxed tissue sections by transmission. Furthermore, the test set was collected with a different brand of FTIR microscope and a different pixel size. Despite the different recording conditions, separating extracellular matrix (ECM) from carcinoma spectra was 100% successful, underlying the robustness of this univariate model and the utility of covariance analysis for revealing efficient wavenumbers. We suggest that 2D covariance maps using the full spectral range could be most useful to select the interesting wavenumbers and achieve very fast data acquisition on quantum cascade laser infrared imaging microscopes.

© 2018 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Breast cancer is the second most common cancer after lung cancer. Its incidence is still growing (by 0.4% during the last 5 years) in the US [1] but increased by 40% among the Gulf Cooperation Council Countries women during the last 12-year period analyzed

* Corresponding author. Center for Structural Biology and Bioinformatics, Laboratory for the Structure and Function of Biological Membranes, Campus Plaine CP206/02, Université Libre de Bruxelles, Bld du Triomphe 2, CP206/2, B1050, Brussels, Belgium.

E-mail address: egoor@ulb.ac.be (E. Goormaghtigh).

[2]. The precise identification of tumor molecular parameters and their integration in cancer management is crucial for improving cancer therapy as well as for moving towards individual therapy. So far, in clinical practice, most cancer parameters originating from histopathology rely on the visualization by a pathologist of microscopic structures observed in stained tissue sections, including immunohistochemistry (IHC) markers (HR, HER2, Ki-67 and/or basal cell markers such as CK5/6 and EGFR) that are largely used in the clinic [4]. In practice, the classifications obtained using these markers help, statistically speaking, define a therapy and provide a prognosis. At individual level however, they are largely insufficient to deal with individual cases. The origin of this failure can be largely

found in the huge heterogeneity of the tumors. Tumor cells mutate at a high rate and a single tumor may contain many different clones [5,6] with different properties which are difficult to identify. A further complexity arises from the interplay between tumor cells and the microenvironment. Finally, the distinct molecular characteristics of the tumors that can have profound implications on clinical behaviors and long-term patient outcomes [7,8] are not fully considered. The result is that most of these parameters, e.g. the histological grade, are associated with inter- and intra-observer discrepancies, in addition to suffering from a quantification problem.

Addressing the heterogeneity of the tumor at microscopic level is not a simple problem. It requires both in depth analysis of the cells contained in the tissue section and a microscopic approach which can provide such an analysis at the cellular level. As far as in-depth analysis is concerned, the development of microarray-based gene expression profiling technologies has helped explain some of this heterogeneity, in particular by sub-classifying human breast carcinomas and establishing molecular-based prognostic and predictive signatures [4,9–12]. The complexity and cost of gene expression profiling limit its use as a routine hospital diagnostic tool. Furthermore, it cannot be applied systematically at the cellular level. This results in an incomplete picture of the heterogeneity of the tumor [13]. Reflecting this cellular heterogeneity therefore requires new approaches.

Some emerging imaging technologies (e.g., mass spectrometry-based “omics” techniques [14,15]) reveal many more biomarkers than immunohistochemical approaches [16,17]. They are promising but not yet robust enough to ensure high-throughput analysis of histological sections in routine pathology. Presently, vibrational spectral histopathology (FTIR or Raman imaging) represents the only method that provides simultaneously, for each pixel of the acquired image, hundreds of robust markers related to the molecular content of the cells [18,19]. Fourier transform infrared (FTIR) spectroscopy has recently shown great potential for disease diagnosis in the field of breast cancer [20–22]. These techniques can provide a biochemical fingerprint of a biopsy sample and, together with advanced data analysis techniques, can accurately classify cell types. Each cellular component has a characteristic set of vibrational transitions resulting in a unique spectrum. It is now considered that vibrational spectroscopies provide as much information as DNA microarrays as far as diagnostic purposes are concerned. Importantly, they can also identify all molecule types as well as details of their chemical structure. For instance the length of lipid acyl chains, the degree of unsaturation of lipids [23], the lipid/protein ratio, DNA condensation state [24] and many other parameters can be obtained from vibrational spectra. In particular, information not only on the chemical nature of cell molecules but also their conformations can be obtained. They are, in particular very sensitive to protein secondary structure [25,26]. Altogether, the various contributions to the spectrum form a signature of the molecular composition of the cell that is unique. Multivariate analysis of spectral data results in remarkably sensitive segmentation of the cells present in a tissue section [27]. Recent progresses in FTIR imaging makes it possible to record images of tissue sections with a spatial resolution close to the cell size. Yet, one of the challenges when dealing with FTIR imaging is the still slow recording and handling of data [19]. When recording FTIR images with a pixel size of $2.7 \times 2.7 \mu\text{m}^2$ as described in this work, a 1 cm^2 image contains 13.4 million spectra which are recorded between 4000 and 900 cm^{-1} for most Focal Plane Array (FPA) detectors, each one containing about 1500 data points. Recording such an image takes several hours, limiting the application of the technology. Yet, not all of this data is useful for identifying cell types. Recording FTIR images with quantum cascade lasers [28–30] and their analysis

would be considerably faster if the useful spectral region were known. For this purpose, analysis of variance is a key issue. Spectral regions with no variance throughout the dataset are obviously of no interest and the co-varying wavenumbers could be reduced to a limited number of wavenumbers as they describe the same variations among cell types.

We propose in the present paper to use 2D correlation analysis to simplify the dataset and possibly identify the few wavenumbers that are of interest for the identification of the various cell types such as carcinoma cells, erythrocytes, lymphocytes and the extracellular matrix (ECM). In the near future, this first approach will have to be followed by an attempt to identify various types of tumor clones in a tissue section.

2. Materials and methods

Spectra of breast carcinoma cells, erythrocytes, lymphocytes and extracellular matrix were obtained from a database maintained in-house. A full description of the samples can be found in Benard et al. [27]. Briefly formalin-fixed paraffin-embedded (FFPE) breast cancer tissues from 66 patients were provided from the tumor tissue bank of the Jules Bordet Institute (Brussels, Belgium). To enrich the database on the immune and stromal responses, seven lymph nodes and three tonsils as well as seven scars from mastectomy biopsied tissue samples were also included. As described in the previous paper [27], more than 13,000 representative spectra were extracted from the images under the supervision of a trained pathologist.

As described previously [27], the FTIR data were collected using a Hyperion 3000 FTIR imaging system (Bruker Optics, Ettlingen, Germany), equipped with a 64×64 Mercury Cadmium Telluride (MCT) Focal Plane Array (FPA) detector. Data were collected in transmission mode from sample regions of $184 \times 184 \mu\text{m}^2$. Each pixel corresponds to an area of $2.7 \times 2.7 \mu\text{m}^2$. One FTIR image (unit image) resulted in 4096 spectra, each one being the average of 256 scans recorded in a spectral range from 3900 to 800 cm^{-1} . The spectral resolution was set to 8 cm^{-1} and data points encoded every 1 cm^{-1} . Spectral processing was performed as described [27]. Briefly, water vapor contribution was subtracted with 1956 – 1935 cm^{-1} as the reference peak. In order to eliminate any intensity variation caused by changes in the thickness of the tissue section or quantity of cellular material, the spectra were normalized for equal area between 1725 and 1481 cm^{-1} . An 11-point baseline correction was subtracted. For this purpose, straight lines were interpolated between the spectral points at 3620 , 2995 , 2800 , 2395 , 2247 , 1765 , 1724 , 1480 , 1355 , 1144 and 950 cm^{-1} and subtracted from each spectrum. Pre-processed spectra were retained for further analyses when the Signal-to-Noise ratio (S/N) was greater than 300:1 for the Amide I and II region (from 1750 to 1480 cm^{-1}). This ratio was calculated considering Signal as the maximum absorbance within 1750 – 1480 cm^{-1} spectra range and Noise as the standard deviation within 2200 – 2100 cm^{-1} range.

For validation, a new set of samples was obtained from the Cancer Registry Office in Al-Amal Hospital, Hamad Medical Corporation (HMC), Doha, Qatar, cut as $5 \mu\text{m}$ thick sections and imaged without removing paraffin. FTIR data were collected using an Agilent 128×128 focal plane array (FPA) mid-IR imager. No binning was applied. Spectra were collected between 3950 and 900 cm^{-1} at a nominal resolution of 8 cm^{-1} and encoded every 2 cm^{-1} . Each spectrum was the mean of 64 scans. The microscope was equipped with a liquid nitrogen cooled 128×128 Mercury Cadmium Telluride (MCT) Focal Plane Array (FPA) detector and a $15 \times$ objective (NA = 0.62). Each pixel corresponds to an area $5.5 \times 5.5 \mu\text{m}^2$. The data were collected in transflection mode from sample regions of $700 \times 700 \mu\text{m}^2$. One FTIR image (unit image) resulted in 16,384

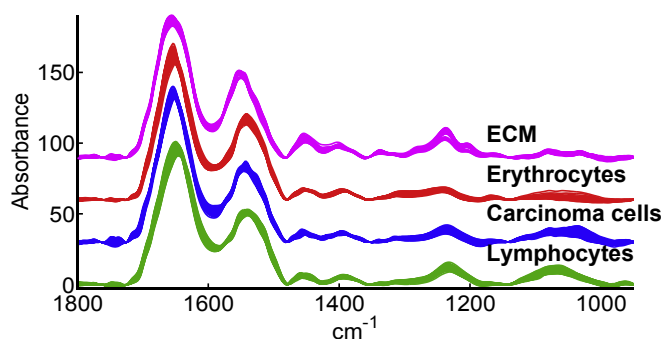


Fig. 1. Overlay of 100 spectra of lymphocytes (green), breast carcinoma cells (blue), erythrocytes (red) and extracellular matrix (ECM) (magenta). Spectra have been processed and normalized as described in Methods. For the sake of clarity, spectra of each cell type have been offset along the absorbance axis. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

spectra. To cover larger sample areas an automatic tiling combined several FTIR unit images in order to obtain one large mosaic FTIR image. This project was approved by the ethics committee.

Two-dimensional (2D) covariance spectra were calculated as described by Noda [31–33]. Computation was carried out using the Hilbert transform, as described [34,35]. Decision trees were computed with Matlab. For building the decision trees, an optimal solution was searched to divide the data set into predefined groups on the basis of a threshold absorbance value at a limited number of wavenumbers. To predict a response, the decisions described in the tree from the root (beginning) node down to a leaf node were followed. The leaf node contains the response. For validation, the sum of the misclassification was obtained for each node. In “restitution” evaluation of the classification tree, the whole data set was used for design and for testing, resulting in an optimistic estimate of the error. In “cross-validation” validation, a 10-fold

cross-validation was used to compute the cost at each node. The sample was first partitioned into 10 subsamples, chosen randomly but with roughly equal size and the same class proportions. For each subsample, a tree is built and tested on the remaining data. The information from all subsamples was used to compute the cost for the whole sample.

Correction of the IR spectra for water vapor and atmospheric CO₂ contribution, baseline subtraction, normalization, application of quality filters, 2D covariance maps, principal component analysis (PCA) and decision tree analyses were carried out by Kinetics, a custom-made program running under Matlab (Mathworks, Inc.).

3. Results

A random selection of 100 spectra belonging to 10 different patients for each cell type, breast carcinoma, erythrocytes, lymphocyte and extracellular matrix (ECM), were used to analyze the variance present throughout the dataset. Fig. 1 provides an overview of the spectra used.

Overall, there is a level of variance within each group as well as between groups which can be observed with the naked eye. There are also between-group differences that can be observed, for instance the weaker absorbance of the extracellular matrix (ECM) spectra in the 1070 cm⁻¹ region.

3.1. 2D covariance analysis

Fig. 2 reports a 2D covariance analysis performed on the 400 spectra presented in Fig. 1. Even though the terminology “2D correlation” is usually used in the literature, we prefer here to use “2D covariance” as spectra have not been normalized by the standard deviation, which would result in a true correlation coefficient map.

It can be observed in Fig. 2 that maxima and minima in the synchronous covariance map appear at discrete locations. The dotted line in Fig. 2 placed at 1670 cm⁻¹ crosses 10 extrema of the

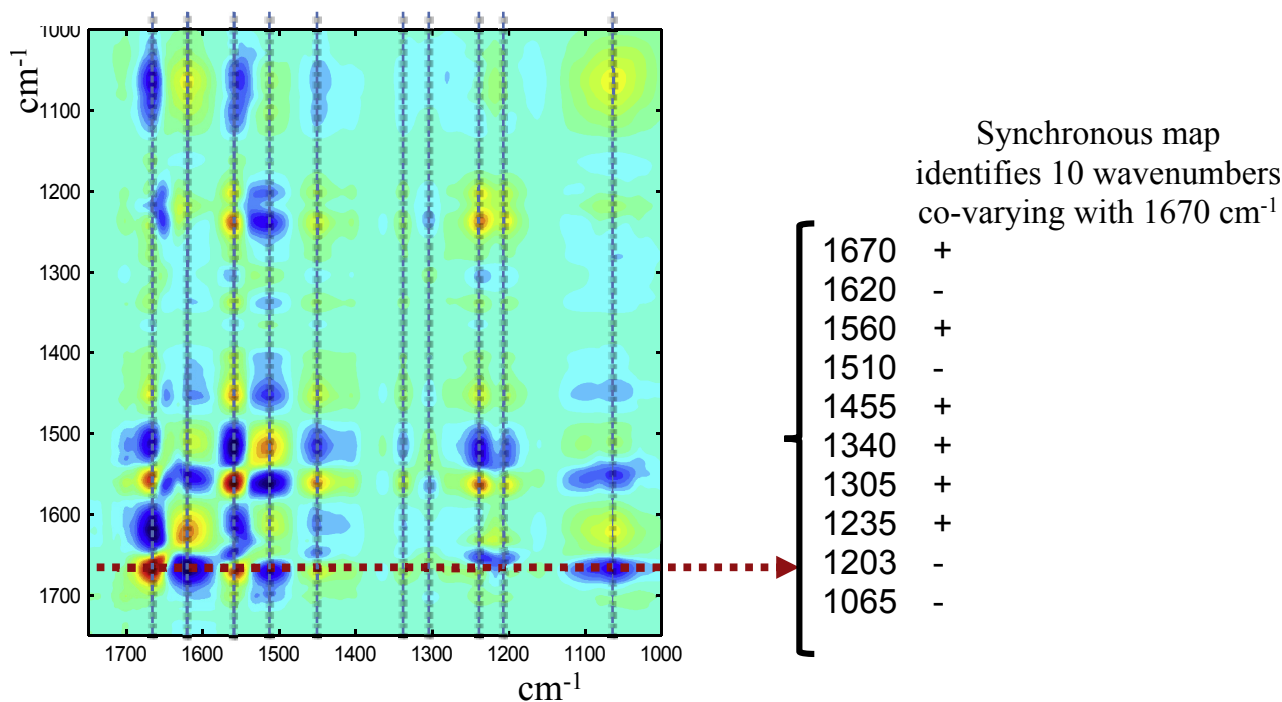


Fig. 2. Synchronous covariance map for the 400 spectra presented in Fig. 1. The red dotted line placed at 1670 cm⁻¹ crosses 10 extrema reported in the right panel. The signs in the right panel indicate whether the covariance is positive or negative. The vertical dotted lines indicate the same wavenumber on the X axis.

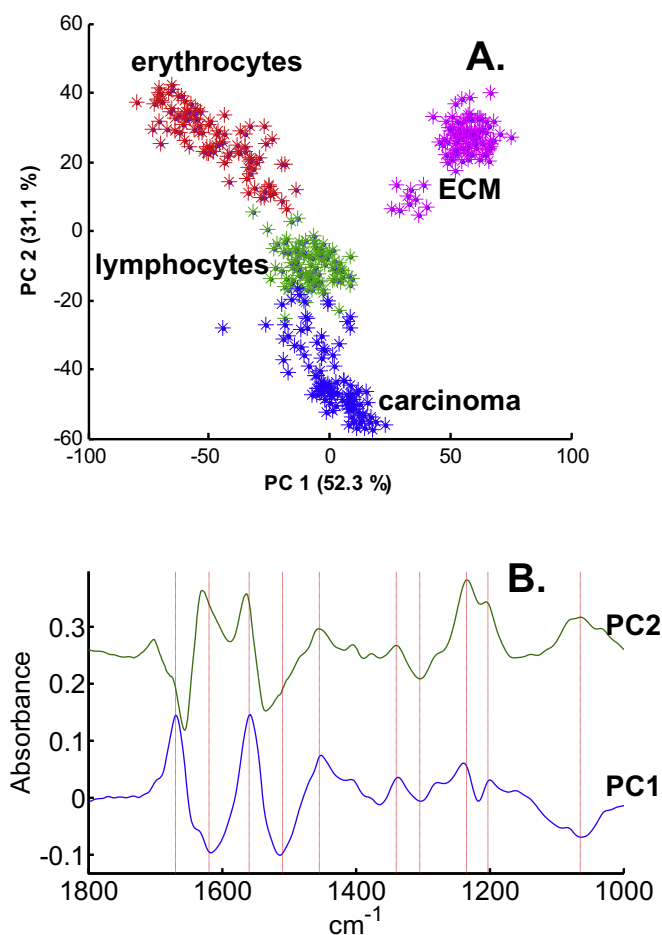


Fig. 3. Principal component analysis of the spectra reported in Fig. 1. A. score plot for 100 spectra of lymphocytes (green), breast carcinoma cells (blue), erythrocytes (red) and extracellular matrix (ECM) (magenta). The fraction of the variance explained by PC1 and PC2 is indicated on the axes, B. the first two principal components. The vertical dotted lines identify the wavenumbers indicated in Fig. 2. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

covariance map that are reported in the right panel with a sign indicating whether the covariance is positive or negative. All covariances found on the diagonal (actually the variances) are obviously positive. It can be observed that the wavenumbers reported in the right panel of Fig. 2 summarize quite well all the wavenumbers presenting extrema in the map. For that reason, it was of interest to examine whether these wavenumbers could be of any use to discriminate the different cell types.

3.2. Principal component analysis (PCA)

PCA is the method of choice to extract co-varying spectral features. It is therefore of interest to compare the synchronous covariance data with the principal components. Fig. 3 reports the result of a principal component analysis performed between 1800 and 1000 cm^{-1} on the spectra presented in Fig. 1. It can be observed that the score plot of the 400 spectra separates well the 4 cell types as shown in Fig. 3A. PC3 further separates lymphocytes from the other classes but further PCs do not contribute significantly to separation (not shown). The wavenumbers identified on Fig. 2 are indicated by the red dotted lines in Fig. 3. There is in general a good match between the synchronous correlation data and PCA, as expected since PCs are the eigen vectors of the covariance matrix.

It can be concluded that PCA identifies the same wavenumbers. One advantage of PCA is that it identifies the wavenumbers that covary (all positive and negative peaks present on the same PC) and the score plot already suggests which PCs are related to the expected discrimination. However, the synchronous map (Fig. 2) might be simpler to read.

3.3. Discrimination based on absorbance at a few wavenumbers

In order to examine the absorbance at each selected wavenumber, the 400 spectra have been represented as images. For instance, Fig. 4 reports the absorbance of the 400 spectra at 1065 cm^{-1} , each absorbance value being represented by a colored pixel of an image. It can be observed that, on the one hand, erythrocytes and ECM form a group and, on the other hand, lymphocytes and carcinoma form another quite distinct group. It therefore appears that the absorbance at 1065 cm^{-1} could be very useful to discriminate between these two groups. This might be rationalized considering the absence of nucleic acid in erythrocytes and the small amounts of nucleic acids in the ECM associated with the presence of only a few scattered cells (essentially fibroblasts). The same process was repeated for the 10 wavenumbers identified by 2D covariance analysis. Results are reported in Fig. 5.

It can now be observed that some wavenumbers almost uniquely identify some specific cell types. In particular, the ECM appears with a much higher absorbance than the other cell types at 1203, 1235, 1340 and 1560 cm^{-1} . The erythrocytes are uniquely identified at 1305 cm^{-1} and the lymphocytes at 1620 cm^{-1} , albeit with a larger degree of confusion.

3.4. Discrimination based on absorbance at a few wavenumbers

In view of the previous observations, we decided to build a decision tree using the absorbances at these 10 wavenumbers to create a model able to discriminate the 4 cell types. Fig. 6 shows such a decision tree.

Interestingly, Fig. 6 reveals that knowledge of the absorbance at only 3 wavenumbers is sufficient to achieve a good classification.

3.5. Discrimination based on absorbance ratios

A potentially more robust approach consists of using ratios of absorbances rather than absorbances. Ratios of absorbances do not depend on scaling, which represents a significant advantage. Yet, one of the problems is the selection of the absorbance ratios that are most discriminant. Considering for instance spectra with $n = 500$ data points, the number of possible combinations is $n \cdot (n-1) / 2$ i.e. 124,750 absorbance ratio values instead of 500 in the original spectrum. Using the 10 wavenumbers identified by 2D covariance analysis (Fig. 2), this number reduces to 45 possible combinations. We rebuilt for each sample "spectra" composed of these 45 absorbance ratios and constructed a decision tree similar to the one shown in Fig. 6. Fig. 7 shows the decision tree obtained using absorbance ratios.

In the training set used, >98% correct classification was achieved for all cell types.

3.6. Validations

A series of 1000 spectra for each cell type were randomly selected among samples which were not included in the previous analysis. Using the rules depicted in the decision tree presented in Fig. 7 and 96.9% of the ECM samples were correctly identified. The figures for correctly identified cell types are 95.1% for carcinoma, 95.1% for erythrocytes and 68% for lymphocytes.

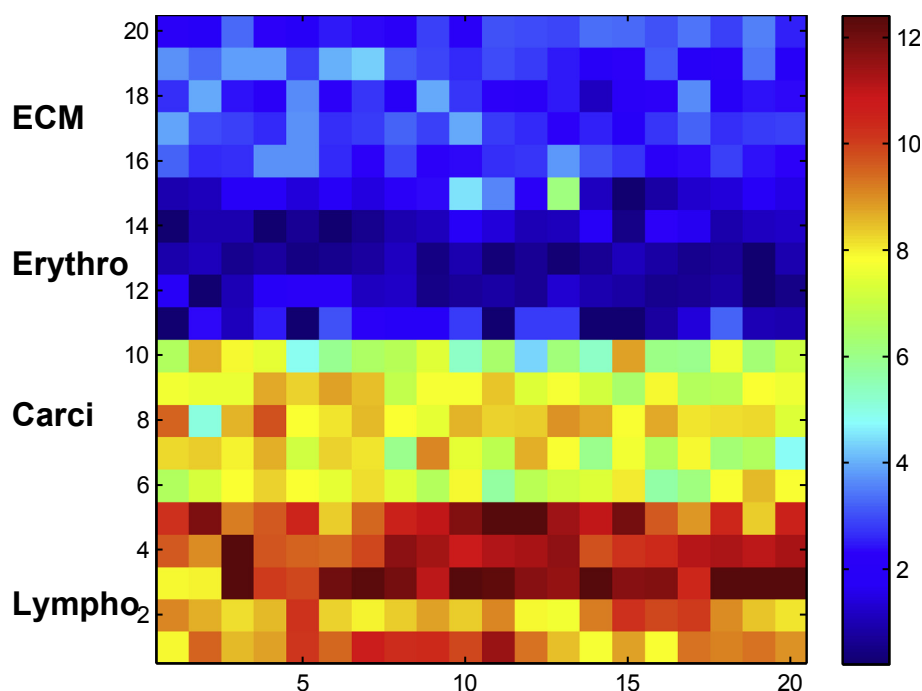


Fig. 4. Absorbance at 1065 cm^{-1} (A^{1065}) of the 400 spectra reported in Fig. 1. Each spectrum is represented as a square, there are 20×20 such squares. The color reports the absorbance according to the color bar on the right hand side of the figure. From the bottom, the first 5 rows report the absorbance of lymphocyte spectra. The next group of 5 rows are A^{1065} carcinoma cells, then of erythrocyte spectra and finally ECM spectra at the top. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

For the sake of generality, we also tested the model on a completely different dataset. For this validation, tissue sections from the Cancer Registry Office in Al-Amal Hospital, Hamad Medical Corporation (HMC), Doha, Qatar, were cut into $5\text{ }\mu\text{m}$ thick sections and FTIR images were recorded with an Agilent 128×128 focal plane array (FPA) mid-IR imager. These tissue sections displayed both carcinoma areas and ECM areas. Here, contrary to the previous samples, paraffin was not removed and spectra were recorded by transfection rather than by transmission resulting in slightly different spectra [36,37]. The advantages of not removing paraffin are: 1) it speeds up the imaging process, 2) it alleviates the scattering problem related to refractive index variations in the sample as it preserves a rather continuous value of the refractive index throughout the image [38] and [3]) it stabilizes the sample by maintaining it in a paraffin environment rather than exposed to air and humidity. Furthermore, paraffin contributions (C-H stretching, $3000\text{--}2800\text{ cm}^{-1}$, and C-H bending, $1500\text{--}1350\text{ cm}^{-1}$) can be discarded from further analyses. Specific algorithms [39–41] have been designed for subtracting paraffin. In our study, contribution of paraffin was merely subtracted to obtain a zero area between 1490 and 1428 cm^{-1} , and this region ($1500\text{--}1400\text{ cm}^{-1}$) was not used for further analyses. In the example shown in Fig. 8, 400 spectra from carcinoma areas and 400 spectra from ECM areas were compared. As the region between 1500 and 1400 cm^{-1} was not used because of uncertainty on the quality of paraffin contribution subtraction, we used the absorbance ratio A^{1670}/A^{1235} . The decision tree reported in Fig. 7 indicated that this ratio should separate carcinoma cells from lymphocytes. Fig. 8 shows perfect discrimination obtained on this completely different sample set, underlying the robustness of the approach. It can be observed that not a single blue pixel can be found in the carcinoma area of the image and not a single yellow-red pixel can be found in the ECM area of the image.

4. Discussion

Our study showed that 2D covariance analysis singles out only a few wavenumbers where both variance and covariance are large. From the 10 selected wavenumbers, simple models could be built to identify the main cell types present in breast cancer tissue sections. Three wavenumbers were necessary to identify correctly 100% of the cell species present in the training test using absorbance values but 4 wavenumbers were necessary when using absorbance ratios. In fact, in the former case, the normalization step applied would require at least one additional wavenumber. It must be noted here that the spectral barcodes proposed by Nallala et al. [42] is another approach that identifies important wavenumbers but it is based on a supervised selection of discriminant spectral ranges characterized by a statistical test and a P-value.

The decision trees provide particularly simple models to reach discrimination between the 4 cell types. It could be expected that a model based on only 4 wavenumbers is less robust than a fully multivariate model. We tested this robustness on a simple model containing only carcinoma and ECM areas. Yet, the challenge was that these samples were obtained in completely different conditions. The sample was recorded by transfection on tissue sections in the presence of paraffin while the training set was obtained on dewaxed tissue sections by transmission. Furthermore, the test set was collected with a different brand of FTIR microscope and a different pixel size. Yet, separating ECM from carcinoma spectra was 100% successful as shown in Fig. 8, underlying the robustness of this univariate model and the utility of covariance analysis for selecting wavenumbers. It must be stressed that in the general case, the contribution of paraffin to the spectra must be carefully considered and approximate subtraction could lead to artifacts if the selected wavenumbers were in a region of strong paraffin absorbance, essentially the $\delta(\text{CH}_2)$ band around 1455 cm^{-1} .

FTIR images of tissues sections obtained with a FPA contain

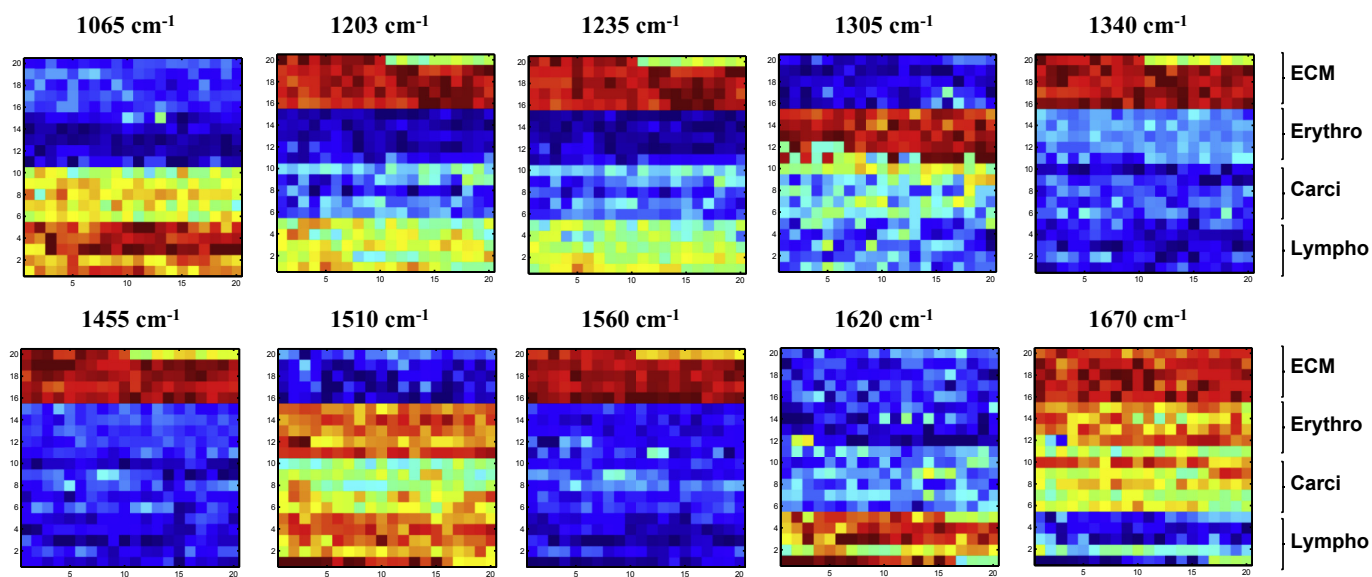


Fig. 5. Absorbance at the wavenumber indicated on top of each image for the 400 spectra reported in Fig. 1. Each spectrum is represented as a square, there are 20 × 20 such squares per image. The color reports the relative absorbance from blue (low) to red (high). As described for Fig. 4, each image contains 5 rows of 20 absorbances from (from bottom to top) lymphocyte spectra, carcinoma cells, erythrocytes and ECM as indicated in the right margin.(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

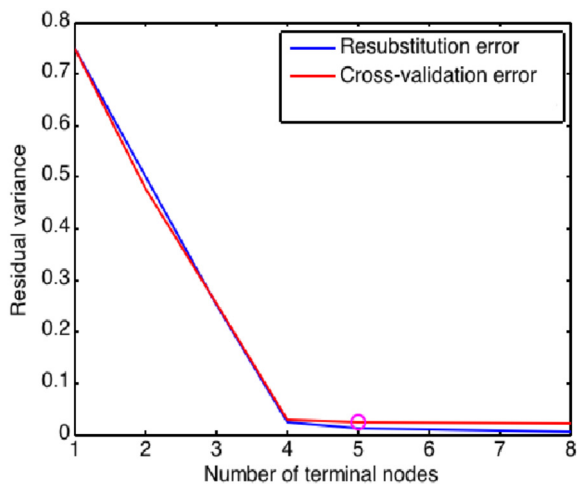
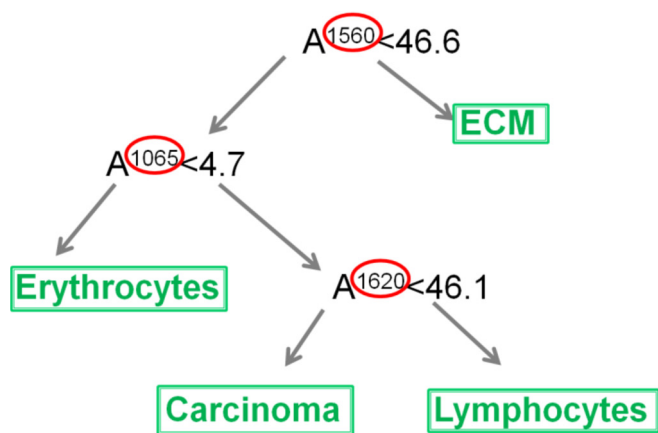


Fig. 6. Top: decision tree built on the basis of the absorbances at the 10 wavenumbers identified by 2D covariance analysis (Fig. 2). Bottom: evolution of the residual variance as a function of the number of terminal nodes obtained by re-substitution (blue line) and cross-validation (red line), see Methods.(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

millions of spectra obtained after several hours of recording, which is too slow for clinical applications. Recent studies on prostate cancer indicate that once spectral biomarkers have been pre-determined, good diagnostics can be obtained based on a limited

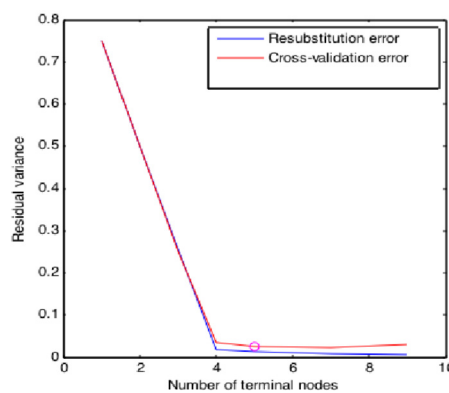
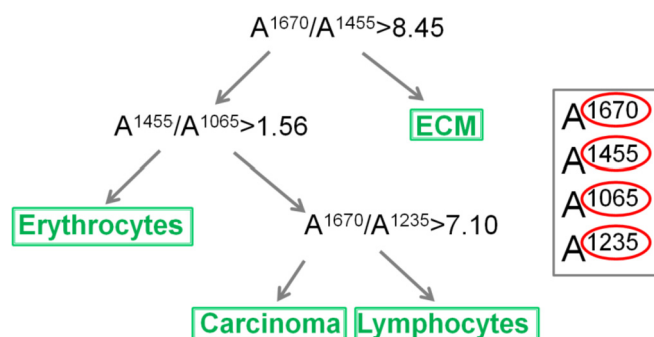


Fig. 7. Top: decision tree built on the basis of the absorbance ratios using the 10 wavenumbers identified by 2D covariance analysis (Fig. 2). The 4 wavenumbers used are circled in red on the right hand side of the figure. Bottom: evolution of the residual variance as a function of the number of terminal nodes obtained by re-substitution (blue line) and cross-validation (red line), see Methods.(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

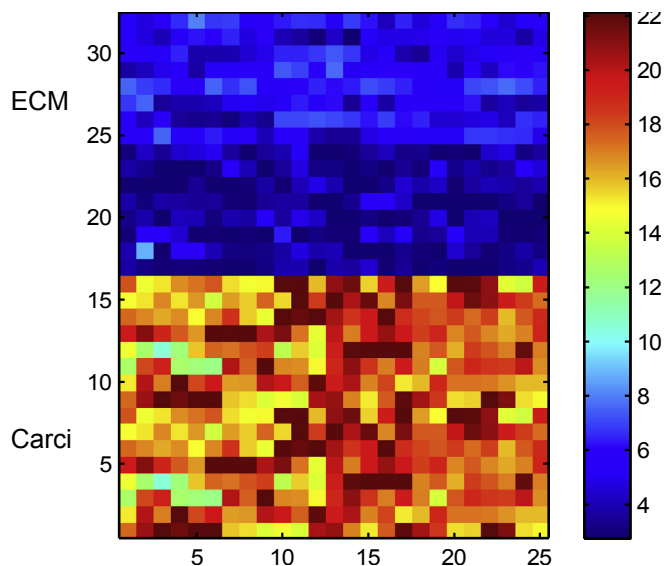


Fig. 8. Absorbance ratio A^{1670}/A^{1235} for 400 spectra of carcinoma and 400 spectra of ECM. Each spectrum is represented as a square. There are 25×16 such squares for carcinoma spectra (bottom) and 25×16 squares for spectra assigned to ECM. The color reports the absorbance ratios according to the color bar on the right hand of the figure. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

number of discrete spectral features [43]. With the advent of broadly tunable quantum cascade lasers (QCL), a new type of FTIR imaging is emerging, resulting in fast discrete frequency IR (DF-IR) spectral imaging [44]. For instance 12 million sparse-frequency spectra (ca 2 cm^2 tissue sample) could be recorded in ca 15 min on a liver biopsy [45]. As reported by Kimber and Kazarian [46], less than 1 min is necessary to record 230,000 spectra counting 9 discrete frequencies necessary to achieve good classification of serum samples [47]. In this context, 2D covariance analysis using the full spectral range could be most useful to select the interesting wavenumbers and achieve very fast data acquisition on QCL-based systems.

Acknowledgements

This work was made possible by a NPRP Award [7 - 1267 - 3–328] from the Qatar National Research Fund (a member of The Qatar Foundation). E.G. is Research Director with the National Fund for Scientific Research (Belgium). The statements made herein are solely the responsibility of the authors.

References

- [1] A. Jemal, E.M. Ward, C.J. Johnson, K.A. Cronin, J. Ma, B. Ryerson, A. Mariotto, A.J. Lake, R. Wilson, R.L. Sherman, R.N. Anderson, S.J. Henley, B.A. Kohler, L. Penberthy, E.J. Feuer, H.K. Weir, Annual report to the nation on the status of cancer, 1975–2014, featuring survival, *J. Natl. Canc. Inst.* 109 (2017), <https://doi.org/10.1093/jnci/djx030>.
- [2] S. Al-Othman, A. Haoudi, S. Alhomoud, A. Alkhenizan, T. Khoja, A. Al-Zahrani, Tackling cancer control in the Gulf cooperation Council Countries, *Lancet Oncol.* 16 (2015) e246–e257, [https://doi.org/10.1016/S1470-2045\(15\)70034-3](https://doi.org/10.1016/S1470-2045(15)70034-3).
- [3] J. Ferlay, H.R. Shin, F. Bray, D. Forman, C. Mathers, D.M. Parkin, *Cancer incidence and mortality worldwide*, in: GLOBOCAN (2008) (Ed.), *Cancer Incid. Mortal. Worldw. IARC CancerBase No. 10* [Internet], v1.2, Lyon, France, 2008.
- [4] A. Prat, C.M. Perou, Deconstructing the molecular portraits of breast cancer, *Mol. Oncol.* 5 (2011) 5–23, <https://doi.org/10.1016/j.molonc.2010.11.003>.
- [5] A.S. Cleary, T.L. Leonard, S.A. Gestl, E.J. Gunther, Tumour cell heterogeneity maintained by cooperating subclones in Wnt-driven mammary cancers, *Nature* 508 (2014) 113–117, <https://doi.org/10.1038/nature13187>.
- [6] L.R. Yates, M. Gerstung, S. Knappskog, C. Desmedt, G. Gundem, P. Van Loo, T. Aas, L.B. Alexandrov, D. Larsson, H. Davies, Y. Li, Y.S. Ju, M. Ramakrishna,

- H.K. Haugland, P.K. Lilleng, S. Nik-Zainal, S. McLaren, A. Butler, S. Martin, D. Glodzik, A. Menzies, K. Raine, J. Hinton, D. Jones, L.J. Mudie, B. Jiang, D. Vincent, A. Greene-Colozzi, P.-Y. Adnet, A. Fatima, M. Maetens, M. Ignatiadis, M.R. Stratton, C. Sotiriou, A.L. Richardson, P.E. Lønning, D.C. Wedge, P.J. Campbell, Subclonal diversification of primary breast cancer revealed by multiregion sequencing, *Nat. Med.* 21 (2015) 751–759, <https://doi.org/10.1038/nm.3886>.
- [7] W.B. Coleman, C.K. Anders, Discerning clinical responses in breast cancer based on molecular signatures, *Am. J. Pathol.* 187 (2017) 2199–2207, <https://doi.org/10.1016/j.ajpath.2017.08.002>.
- [8] J.A. Joyce, J.W. Pollard, Microenvironmental regulation of metastasis, *Nat. Rev. Canc.* 9 (2009) 239–252, <https://doi.org/10.1038/nrc2618>.
- [9] C.M. Perou, T. Sørlie, M.B. Eisen, M. van de Rijn, S.S. Jeffrey, C.A. Rees, J.R. Pollack, D.T. Ross, H. Johnsen, L.A. Akslen, Ø. Fluge, A. Pergamenschikov, C. Williams, S.X. Zhu, P.E. Lønning, A.-L. Børresen-Dale, P.O. Brown, D. Botstein, Molecular portraits of human breast tumours, *Nature* 406 (2000) 747–752, <https://doi.org/10.1038/35021093>.
- [10] T. Sørlie, C.M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M.B. Eisen, M. van de Rijn, S.S. Jeffrey, T. Thorsen, H. Quist, J.C. Matese, P.O. Brown, D. Botstein, P.E. Lønning, A.-L. Børresen-Dale, Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications, *Proc. Natl. Acad. Sci.* 98 (2001) 10869–10874, <https://doi.org/10.1073/pnas.191367098>.
- [11] B. Weigelt, F.L. Baehner, J.S. Reis-Filho, The contribution of gene expression profiling to breast cancer classification, prognosis and prediction: a retrospective of the last decade, *J. Pathol.* 220 (2009) 263–280, <https://doi.org/10.1002/path.2648>.
- [12] J.S. Parker, M. Mullins, M.C.U. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, J.F. Quackenbush, I.J. Stijleman, J. Palazzo, J.S. Marron, A.B. Nobel, E. Mardis, T.O. Nielsen, M.J. Ellis, C.M. Perou, P.S. Bernard, Supervised risk predictor of breast cancer based on intrinsic subtypes, *J. Clin. Oncol.* 27 (2009) 1160–1167, <https://doi.org/10.1200/JCO.2008.18.1370>.
- [13] N. Navin, J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D. Levy, D. Esposito, L. Muthuswamy, A. Krasnitz, W.R. McCombie, J. Hicks, M. Wigler, Tumour evolution inferred by single-cell sequencing, *Nature* 472 (2011) 90–94, <https://doi.org/10.1038/nature09807>.
- [14] A.K. Casasent, M. Edgerton, N.E. Navin, Genome evolution in ductal carcinoma in situ: invasion of the clones, *J. Pathol.* 241 (2017) 208–218, <https://doi.org/10.1002/path.4840>.
- [15] D. Alberts, C. Pottier, N. Smargiasso, D. Baiwir, G. Mazzucchelli, P. Delvenne, M. Kriegsmann, D. Kazdal, A. Warth, E. De Pauw, R. Longuespée, MALDI imaging-guided microproteomic analyses of heterogeneous breast tumors—a pilot study, *Proteomics Clin. Appl.* 12 (2018) 1700062, <https://doi.org/10.1002/prca.201700062>.
- [16] C.N. Ferguson, J.W.M. Fowler, J.F. Waxer, R.A. Gatti, J.A. Loo, Mass spectrometry-based tissue imaging of small molecules, *Adv. Exp. Med. Biol.* 806 (2014) 283–299, https://doi.org/10.1007/978-3-319-06068-2_12.
- [17] J. Kriegsmann, M. Kriegsmann, R. Casadonte, MALDI TOF imaging mass spectrometry in clinical pathology: a valuable tool for cancer diagnostics (Review), *Int. J. Oncol.* 46 (2015) 893–906, <https://doi.org/10.3892/ijo.2014.2788>.
- [18] M.J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H.J. Butler, K.M. Dorling, P.R. Fielden, S.W. Fogarty, N.J. Fullwood, K.A. Heys, C. Hughes, P. Lasch, P.L. Martin-Hirsch, B. Obinaju, G.D. Sockalingum, J. Sulé-Suso, R.J. Strong, M.J. Walsh, B.R. Wood, P. Gardner, F.L. Martin, Using Fourier transform IR spectroscopy to analyze biological materials, *Nat. Protoc.* 9 (2014) 1771–1791, <https://doi.org/10.1038/nprot.2014.110>.
- [19] E. Goormaghtigh, *Infrared imaging in histopathology: is a unified approach possible?* *Biomed. Spectrosc. Imag.* 5 (2016) 325–346.
- [20] S.E. Holton, A. Bergamaschi, B.S. Katzenellenbogen, R. Bhargava, Integration of molecular profiling and chemical imaging to elucidate fibroblast-microenvironment impact on cancer cell phenotype and endocrine resistance in breast cancer, *PLoS One* 9 (2014) e96878, <https://doi.org/10.1371/journal.pone.0096878>.
- [21] M. Verdonck, A. Denayer, B. Delvaux, S. Garaud, R. De Wind, C. Desmedt, C. Sotiriou, K. Willard-Gallo, E. Goormaghtigh, Characterization of human breast cancer tissues by infrared imaging, *Analyst* 141 (2016) 606–619, <https://doi.org/10.1039/c5an01512j>.
- [22] S. Kumar, C. Desmedt, D. Larsson, C. Sotiriou, E. Goormaghtigh, Change in the microenvironment of breast cancer studied by FTIR imaging, *Analyst* 138 (2013) 4058–4065, <https://doi.org/10.1039/c3an00241a>.
- [23] A. Derenne, T. Claessens, C. Conus, E. Goormaghtigh, *Infrared spectroscopy of membrane lipids*, *Encycl. Biophys.* (2013) 1074–1081.
- [24] D.R. Whelan, K.R. Bamberg, P. Heraud, M.J. Tobin, M. Diem, D. McNaughton, B.R. Wood, Monitoring the reversible B to A-like transition of DNA in eukaryotic cells using Fourier transform infrared spectroscopy, *Nucleic Acids Res.* 39 (2011) 5439–5448, <https://doi.org/10.1093/nar/gkr175>.
- [25] E. Goormaghtigh, R. Gasper, A. Benard, A. Goldsztein, V. Raussens, A. Benard, Protein secondary structure content in solution, films and tissues: redundancy and complementarity of the information content in circular dichroism, transmission and ATR FTIR spectra, *Biochim. Biophys. Acta-Proteins Proteomics* 1794 (2009) 1332–1343, <https://doi.org/10.1016/j.bbapap.2009.06.007>.
- [26] E. Goormaghtigh, J.M. Ruyschaert, V. Raussens, Evaluation of the information content in infrared spectra for protein secondary structure determination, *Biophys. J.* 90 (2006) 2946–2957, <https://doi.org/10.1523/JNEUROSCI.0002.2006>.

- [27] A. Benard, C. Desmedt, M. Smolina, P. Szternfeld, M. Verdonck, G. Rouas, N. Kheddoumi, F. Rothé, D. Larsimont, C. Sotiriou, E. Goormaghtigh, Infrared imaging in breast cancer: automated tissue component recognition and spectral characterization of breast cancer cells as well as the tumor micro-environment, *Analyst* 139 (2014) 1044–1056, <https://doi.org/10.1039/c3an01454a>.
- [28] B. Bird, M.J. Baker, Quantum cascade lasers in biomedical infrared imaging, *Trends Biotechnol.* 33 (2015) 557–558, <https://doi.org/10.1016/j.tibtech.2015.07.003>.
- [29] R. Bhargava, Infrared spectroscopic imaging: the next generation, *Appl. Spectrosc.* 66 (2012) 1091–1120, <https://doi.org/10.1366/12-06801>.
- [30] M.J. Pilling, A. Henderson, B. Bird, M.D. Brown, N.W. Clarke, P. Gardner, High-throughput quantum cascade laser (QCL) spectral histopathology: a practical approach towards clinical translation, *Faraday Discuss* (2016), <https://doi.org/10.1039/c5fd00176e>.
- [31] I. Noda, Generalized 2-dimensional correlation method applicable to infrared, Raman, and other types of spectroscopy, *Appl. Spectrosc.* 47 (1993) 1329–1336 [isi:A1993LX76000006](https://doi.org/10.1366/11-06568).
- [32] I. Noda, 2-Dimensional infrared (2D Ir) spectroscopy - theory and applications, *Appl. Spectrosc.* 44 (1990) 550–561 [isi:A1990DD24000004](https://doi.org/10.1366/11-06568).
- [33] I. Noda, A.E. Dowrey, C. Marcott, G.M. Story, Y. Ozaki, Generalized two-dimensional correlation spectroscopy, *Appl. Spectrosc.* 54 (2000) 236A–248A [isi:000088499100002](https://doi.org/10.1366/11-06568).
- [34] I. Noda, Determination of two-dimensional correlation spectra using the Hilbert transform, *Appl. Spectrosc.* 54 (2000) 994–999 [isi:000088499100012](https://doi.org/10.1366/11-06568).
- [35] S. Sasic, A. Muszynski, Y. Ozaki, New insight into the mathematical background of generalized two-dimensional correlation spectroscopy and the influence of mean normalization pretreatment on two-dimensional correlation spectra, *Appl. Spectrosc.* 55 (2001) 343–349 [isi:000167864800016](https://doi.org/10.1366/11-06568).
- [36] R.K. Reddy, M.J. Walsh, M. V. Schulmerich, P.S. Carney, R. Bhargava, High-definition infrared spectroscopic imaging, *Appl. Spectrosc.* 67 (2013) 93–105, <https://doi.org/10.1366/11-06568>.
- [37] V. Zohdi, D.R. Whelan, B.R. Wood, J.T. Pearson, K.R. Bambery, M.J. Black, Importance of tissue preparation methods in FTIR micro-spectroscopical analysis of biological tissues: “traps for new users”, *PLoS One* 10 (2015) <https://doi.org/10.1371/journal.pone.0116491> e0116491.
- [38] B. Bird, K. Bedrossian, N. Laver, M. Miljković, M.J. Romeo, M. Diem, Detection of breast micro-metastases in axillary lymph nodes by infrared micro-spectral imaging, *Analyst* 134 (2009) 1067–1076, <https://doi.org/10.1039/b821166c>.
- [39] E. Ly, O. Piot, R. Wolthuis, A. Durlach, P. Bernard, M. Manfait, Combination of FTIR spectral imaging and chemometrics for tumour detection from paraffin-embedded biopsies, *Analyst* 133 (2008) 197–205, <https://doi.org/10.1039/b715924b>.
- [40] D. Sebiskveradze, V. Vrabie, C. Gobinet, A. Durlach, P. Bernard, E. Ly, M. Manfait, P. Jeannesson, O. Piot, Automation of an algorithm based on fuzzy clustering for analyzing tumoral heterogeneity in human skin carcinoma tissue sections, *Lab. Invest.* 91 (2011) 799–811, <https://doi.org/10.1038/labinvest.2011.13>.
- [41] T.N.Q. Nguyen, P. Jeannesson, A. Groh, O. Piot, D. Guenot, C. Gobinet, Fully unsupervised inter-individual IR spectral histology of paraffinized tissue sections of normal colon, *J. Biophot.* 9 (2016) 521–532, <https://doi.org/10.1002/jbio.201500285>.
- [42] J. Nallala, O. Piot, M.-D. Diebold, C. Gobinet, O. Bouché, M. Manfait, G.D. Sockalingum, Infrared imaging as a cancer diagnostic tool: introducing a new concept of spectral barcodes for identifying molecular changes in colon tumors, *Cytometry A*. 83 (2013) 294–300, <https://doi.org/10.1002/cyto.a.22249>.
- [43] M.J. Pilling, A. Henderson, B. Bird, M.D. Brown, N.W. Clarke, P. Gardner, High-throughput quantum cascade laser (QCL) spectral histopathology: a practical approach towards clinical translation, *Faraday Discuss* 187 (2016) 135–154, <https://doi.org/10.1039/c5fd00176e>.
- [44] K. Yeh, S. Kenkel, J.-N. Liu, R. Bhargava, Fast infrared chemical imaging with a quantum cascade laser, *Anal. Chem.* 87 (2015) 485–493, <https://doi.org/10.1021/ac5027513>.
- [45] B. Bird, J. Rowlette, A protocol for rapid, label-free histochemical imaging of fibrotic liver, *Analyst* 142 (2017) 1179–1184, <https://doi.org/10.1039/C6AN02080A>.
- [46] J.A. Kimber, S.G. Kazarian, Spectroscopic imaging of biomaterials and biological systems with FTIR microscopy or with quantum cascade lasers, *Anal. Bioanal. Chem.* 409 (2017) 5813–5820, <https://doi.org/10.1007/s00216-017-0574-5>.
- [47] C. Hughes, G. Clemens, B. Bird, T. Dawson, K.M. Ashton, M.D. Jenkinson, A. Brodbelt, M. Weida, E. Fotheringham, M. Barre, J. Rowlette, M.J. Baker, Introducing discrete frequency infrared technology for high-throughput bio-fluid screening, *Sci. Rep.* 6 (2016) 20173, <https://doi.org/10.1038/srep20173>.